ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 12, Iss 1, Jan 2023

A Novel Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs

VEERA SIVA PRASAD[1], VIJAY KUMAR PADALA[2], CHOTAPALLI. KISHORE BABU[3], SURYA VARA PRASAD NEETIPUDI [4]

[1] ASST PROFESSOR DEPARTMENT OF COMPUTER SCIENCE, SIR C R REDDY COLLEGE, ELURU, INDIA.
[2] ASST PROFESSOR DEPARTMENT OF COMPUTER SCIENCE, SIR C R REDDY COLLEGE, ELURU, INDIA.
[3] ASST PROFESSOR DEPARTMENT OF COMPUTER SCIENCE, SIR C R REDDY COLLEGE, ELURU, INDIA.
[4] ASST PROFESSOR DEPARTMENT OF COMPUTER SCIENCE, SIR C R REDDY COLLEGE, ELURU, INDIA.

spv@sircrreddycollege.ac.in[1], pvk@sircrreddycollege.ac.in[2], ckb@sircrreddycollege.ac.in[3], neethipudisvprasad@gmail.com[4]

Abstract

Accurately predicting students' future performance based on their ongoing academic records is crucial for effectively carrying out necessary pedagogical interventions to ensure students' on-time and satisfactory graduation. Although there is a rich literature on predicting student performance when solving problems or studying for courses using data-driven approaches, predicting student performance in completing degrees (e.g. college programs) is much less studied and faces new challenges: (1) Students differ tremendously in terms of backgrounds and selected courses; (2) Courses are not equally informative for making accurate predictions; (3) Students' evolving progress needs to be incorporated into the prediction. In this paper, we develop a novel machine learning method for predicting student performance in degree programs that is able to address these key challenges. The proposed method has two major features. First, a bilayered structure comprising of multiple base predictors and a cascade of ensemble predictors is developed for making predictions based on students' evolving performance states. Second, a data-driven approach based on latent factor models and probabilistic matrix factorization is proposed to discover course relevance, which is important for constructing efficient base predictors. Through extensive simulations on an undergraduate student dataset collected over three years at UCLA, we show that the proposed method achieves superior performance to benchmark approaches.

Keywords: Machine Learning, latent factor models, UCLA

I. INTRODUCTION

Making higher education affordable has a significant impact on ensuring the nation's economic prosperity and represents a central focus of the government when making education policies. Yet student loan debt in the United States has blown past the trillion-dollar mark, exceeding Americans' combined credit card and auto loan debts. As the cost in college education (tuitions, fees and living expenses) has skyrocketed over the past few decades, prolonged graduation time has become a crucial contributing factor to the evergrowing student loan debt. In fact, recent studies show that only 50 of the more than 580 public four-year institutions in the United States have on-time graduation rates at or above 50 percent for their full-time students. To make college more affordable, it is thus crucial to ensure that many more students graduate on time through early interventions on students whose performance will be unlikely to meet the graduation criteria of the degree program on time. A critical step towards effective intervention is to build a system that can continuously keep track of students' academic performance and accurately predict their future performance, such as when they are likely to graduate and their estimated final GPAs, given the current progress. Although predicting student performance has been extensively studied in the literature, it was primarily studied in the contexts of solving problems in Intelligent Tutoring Systems (ITSs) or completing courses in classroom settings or in Massive Open Online Courses (MOOC) platforms. However, predicting student performance within a degree program (e.g. college program) is significantly different and faces new challenges. First, students can differ tremendously in terms of backgrounds as well as their chosen areas (majors, specializations), resulting in different selected courses as well as course sequences. On the other hand, the same course can be taken by students in different areas. Since predicting student performance in a particular course relies on the student past performance in other courses, a key challenge for training an

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 12, Iss 1, Jan 2023

effective predictor is how to handle heterogeneous student data due to different areas and interests. In contrast, solving problems in ITSs often follow routine steps which are the same for all students. Similarly, predictions of students' performance in courses are often based on in-course assessments which are designed to be the same for all students. Second, students may take many courses but not all courses are equally informative for predicting students' future performance. Utilizing the student's past performance in all courses that he/she has completed not only increases complexity but also introduces noise in the prediction, thereby degrading the prediction performance. For instance, while it makes sense to consider a student's grade in the course "Linear Algebra" for predicting his/her grade in the course "Linear Optimization", the student's grade in the course "Chemistry Lab" may have much weaker predictive power. However, the course correlation is not always as obvious as in this case. Therefore, discovering the underlying correlation among courses is of great importance for making accurate performance predictions. Third, predicting student performance in a degree program is not a one-time task; rather, it requires continuous tracking and updating as the student finishes new courses over time. An important consideration in this regard is that the prediction needs to be made based on not only the most recent snapshot of the student accomplishments but also the evolution of the student progress, which may contain valuable information for making more accurate predictions. However, the complexity can easily explode since even mathematically representing the evolution of student progress itself can be a daunting task. However, treating the past progress equally as the current performance when predicting the future may not be a wise choice either since intuition tells us that old information tends to be outdated. In light of the aforementioned challenges, in this paper, we propose a novel method for predicting student performance in a degree program. We focus on predicting students' GPAs but the general framework can be used for other student performance prediction tasks.

II. METHODOLOGY

EXISTING SYSTEM

In fact, recent studies show that only 50 of the more than 580 public four-year institutions in the United States have on-time graduation rates at or above 50 percent for their full-time students . To make college more affordable, it is thus crucial to ensure that many more students graduate on time through early interventions on students whose performance will be unlikely to meet the graduation criteria of the degree program on time. A critical step towards effective intervention is to build a system that can continuously keep track of students' academic performance and accurately predict their future performance, such as when they are likely to graduate and their estimated final GPAs, given the current progress. Although predicting student performance has been extensively studied in the literature, it was primarily studied in the contexts of solving problems in Intelligent Tutoring Systems (ITSs).

DISADVANTAGES OF EXISTING SYSTEM

However, predicting student performance within a degree program (e.g. college program) is significantly different and faces new challenges.

PROPOSED SYSTEM

We consider a degree program in which students must complete a set of courses to graduate in T academic terms. Courses have prerequisite dependencies, namely a course can be taken only when certain prerequisite courses have been taken and passed. In general, the prerequisite dependency can be described as a directed acyclic graph (DAG) . There can be multiple specialized areas in a program which require different subsets of courses to be completed for students to graduate. We will focus on the prediction problem for one area in this department. Nevertheless, data from other areas will still be utilized for our prediction tasks. The reason is that data from a single area is often limited while different areas still share many common courses.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 12, Iss 1, Jan 2023

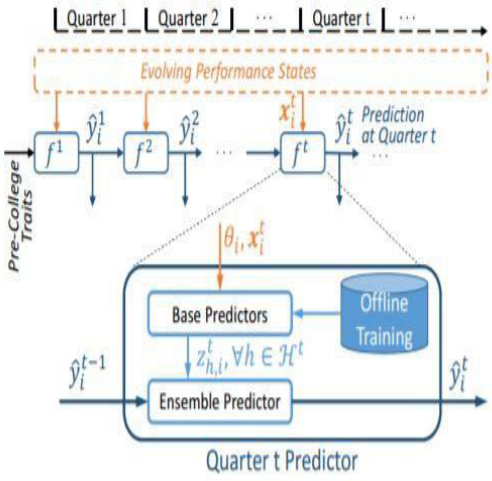

Figure 1: Proposed architecture

ADVANTAGES OF PROPOSED SYSTEM

- It is important for constructing efficient base predictors.
- System that can continuously keep track of students' academic performance and accurately predict their future performance.

SYSTEM REQUIREMENTS SPECIFICATIONS:

The functional requirements or the overall description documents include the product perspective and features, operating system and operating environment, graphics requirements, design constraints and user documentation.

The appropriation of requirements and implementation constraints gives the general overview of the project in regard to what the areas of strength and deficit are and how to tackle them.

- ✓ Python idel 3.7 version (or)
- ✓ Anaconda 3.7 (or)
- ✓ Jupiter (or)
- ✓ Google colab

III. RESULTS & DISCUSSION

To run this project double click on 'run.bat' file to get below screen

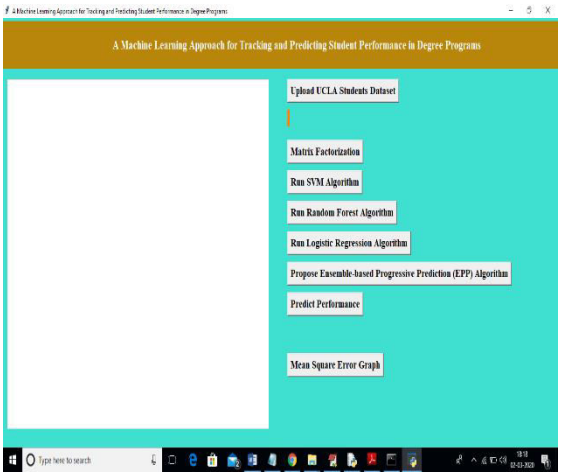

Figure 2: Upload UCLA Students Dataset

In above screen click on 'Upload UCLA Students Dataset' button to upload dataset

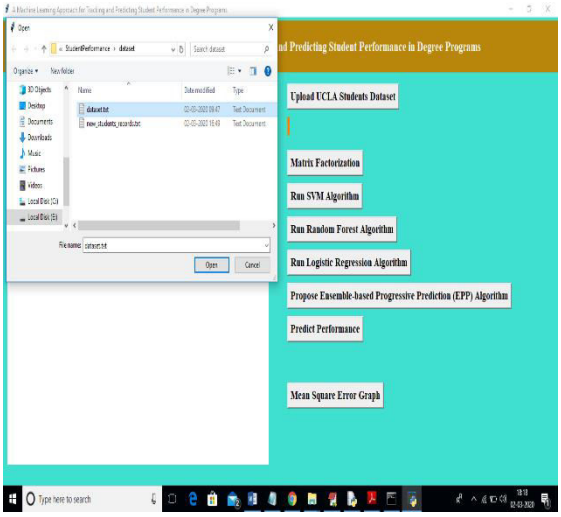

Figure 3: uploading 'dataset.txt''

In above screen I am uploading 'dataset.txt'' as student dataset. After uploading will get below screen

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 12, Iss 1, Jan 2023

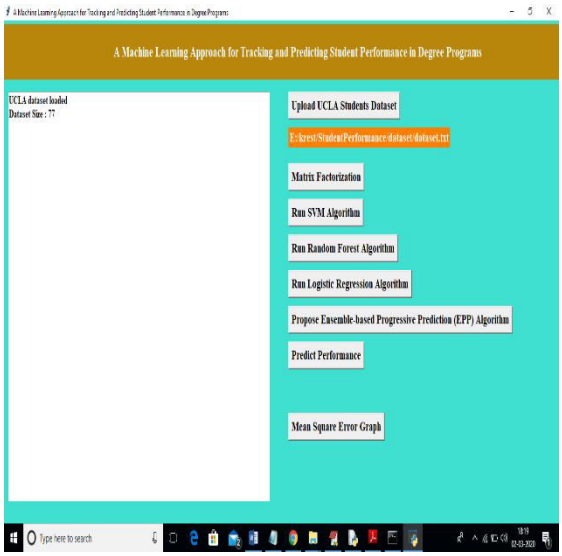

Figure 4: dataset contains total 77 student's records.

In above screen we can see dataset contains total 77 student's records. Now click on 'Matrix Factorization' to build feature vector from dataset. In this matrix we will have all related course data and if student taken course then matrix contains marks otherwise 0.

Figure 5: all records converted to feature vector.

In above screen we can see all records converted to feature vector and in above screen in first 3 lines we can see from above matrix application using 61 records to train machine learning model and 16 records to test accuracy or to calculate Mean Square Error of classifier. If algorithm prediction result is high then accuracy will be more and Mean Square Error (MSE) will be less. Now we got matrix and data to train and test classifier. Now click on 'Run SVM Algorithm' to train SVM classifier and to get it accuracy and MSE value

Figure 6: SVM MSE is 56%.

In above screen SVM MSE is 56%. Now click on 'Run Random Forest Algorithm' to generate training model using Random Forest and to get it accuracy and MSE

Figure 7: random forest got 43% MSE.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 12, Iss 1, Jan 2023

In above screen random forest got 43% MSE and now click on 'Run Logistic Regression Algorithm' button to get it accuracy and MSE

Figure 8: screen logistic regression got 43% MSE.

In above screen logistic regression got 43% MSE and now click on 'Propose Ensemble-based Progressive Prediction (EPP) Algorithm' button to generate model using propose EPP algorithm and to get it accuracy and MSE

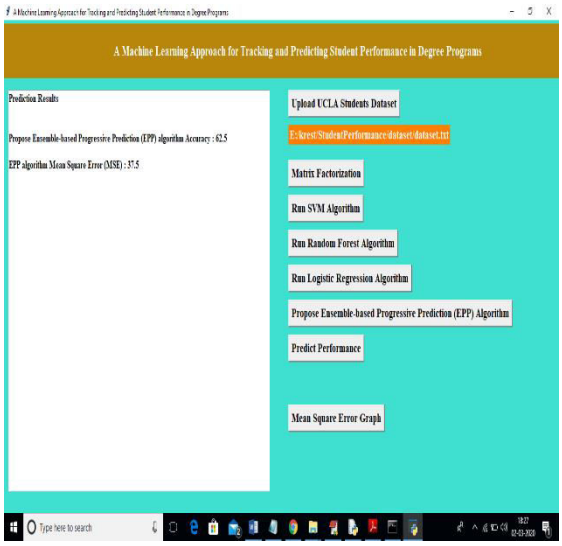

Figure 9: EPP propose algorithm got 37% MSE.

In above screen EPP propose algorithm got 37% MSE and now click on 'Predict Performance' button to upload student on going test marks and to predict GPA for future course

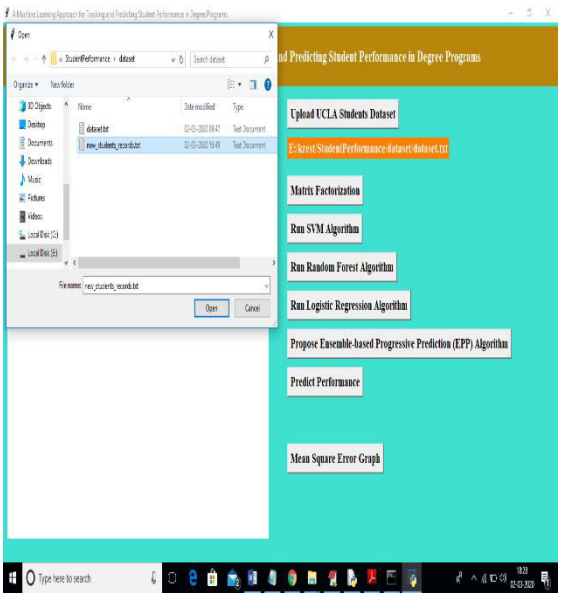

Figure 10: uploading new student records as test file.

In above screen uploading new student records as test file and below are the prediction results

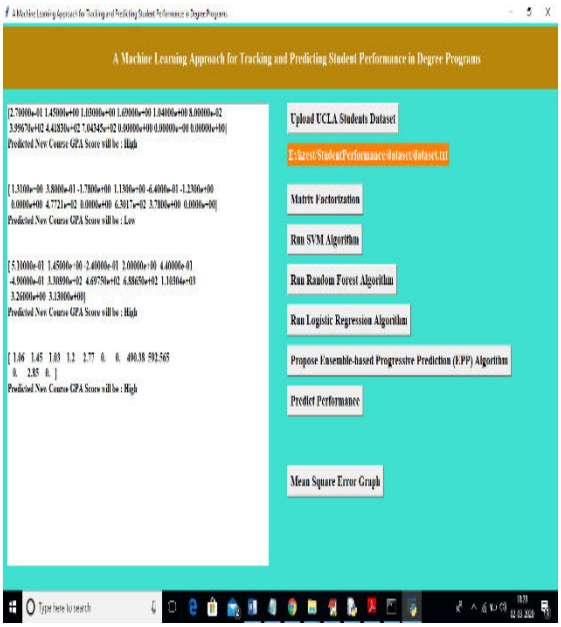

Figure 11: screen in square brackets are the student marks

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 12, Iss 1, Jan 2023

In above screen in square brackets are the student marks of ongoing subjects and this marks are converted to matrix factorization and then applied on EPP train model to predict GPA as LOW or HIGH. In above screen after each test record I am displaying predicted result value. Now click on "Mean Square Error Graph' button to get below graph

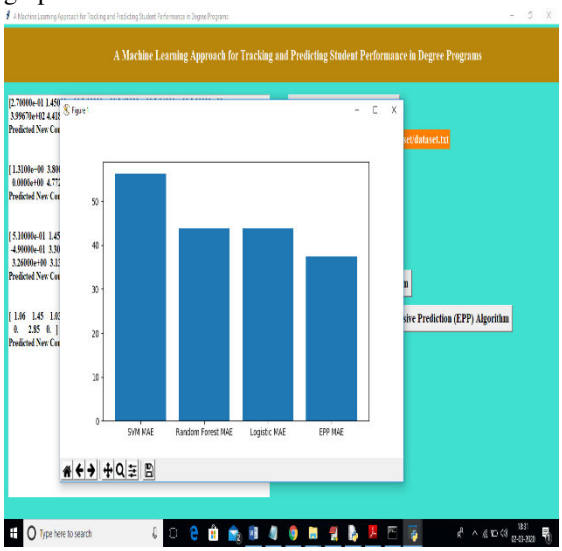

Figure 12: x-axis represents algorithm name and y-axis represents MSE.

In above graph x-axis represents algorithm name and y-axis represents MSE (mean square error). From above graph we can see propose algorithm got less MSE error and has high accuracy compare to other algorithms. From above graph we can conclude that propose EPP is better in prediction compare to other algorithms

IV. CONCLUSION

In this paper, we proposed a novel method for predicting students' future performance in degree programs given their current and past performance. A latent factor model-based course clustering method was developed to discover relevant courses for constructing base predictors. An ensemble-based progressive prediction architecture was developed to incorporate students' evolving performance into the prediction. These datadriven methods can be used in conjunction with other pedagogical methods for evaluating students' performance and provide valuable information for academic advisors to recommend subsequent courses to students and carry out pedagogical intervention measures if necessary. Additionally, this work will also impact curriculum design in degree programs and education policy design in general. Future work includes extending the performance prediction to elective courses and using the prediction results to recommend courses to students.

REFERENCE

[1] The White House, "Making college affordable," https:// www.whitehouse.gov/issues/education/higher-education/ making-college-affordable, 2016.

[2] Complete College America, "Four-year myth: Making college more affordable," http://completecollege.org/wp-content/uploads/2014/ 11/4-Year-Myth.pdf, 2014.

[3] H. Cen, K. Koedinger, and B. Junker, "Learning factors analysis–a general method for cognitive model evaluation and improvement," in International Conference on Intelligent Tutoring Systems. Springer, 2006, pp. 164–175.

[4] M. Feng, N. Heffernan, and K. Koedinger, "Addressing the assessment challenge with an online system that tutors as it assesses," User Modeling and User-Adapted Interaction, vol. 19, no. 3, pp. 243–266, 2009.

[5] H.-F. Yu, H.-Y. Lo, H.-P. Hsieh, J.-K. Lou, T. G. McKenzie, J.-W. Chou, P.-H. Chung, C.-H. Ho, C.-F. Chang, Y.-H. Wei et al., "Feature engineering and classifier ensemble for kdd cup 2010," in Proceedings of the KDD Cup 2010 Workshop, 2010, pp. 1–16.

[6] Z. A. Pardos and N. T. Heffernan, "Using hmms and bagged decision trees to leverage rich features of user and skill from an intelligent tutoring system dataset," Journal of Machine Learning Research W & CP, 2010.

[7] Y. Meier, J. Xu, O. Atan, and M. van der Schaar, "Personalized grade prediction: A data mining approach," in Data Mining (ICDM), 2015 IEEE International Conference on. IEEE, 2015, pp. 907–912.

[8] C. G. Brinton and M. Chiang, "Mooc performance prediction via clickstream data and social learning networks," in 2015 IEEE Conference on Computer Communications (INFOCOM). IEEE, 2015, pp. 2299– 2307.

[9] KDD Cup, "Educational data minding challenge," https://pslcdatashop. web.cmu.edu/KDDCup/, 2010.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 12, Iss 1, Jan 2023

[10] Y. Jiang, R. S. Baker, L. Paquette, M. San Pedro, and N. T. Heffernan, "Learning, moment-by-moment and over the long term," in International Conference on Artificial Intelligence in Education. Springer, 2015, pp. 654–657.

[11] C. Marquez-Vera, C. Romero, and S. Ventura, "Predicting school failure using data mining," in Educational Data Mining 2011, 2010.

[12] Y.-h. Wang and H.-C. Liao, "Data mining for adaptive learning in a tesl-based e-learning system," Expert Systems with Applications, vol. 38, no. 6, pp. 6480–6485, 2011.

[13] N. Thai-Nghe, L. Drumond, T. Horvath, L. Schmidt-Thieme ´ et al., "Multi-relational factorization models for predicting student performance," in Proc. of the KDD Workshop on Knowledge Discovery in Educational Data. Citeseer, 2011.

[14] A. Toscher and M. Jahrer, "Collaborative filtering applied to educational data mining," KDD cup, 2010.

[15] R. Bekele and W. Menzel, "A bayesian approach to predict performance of a student (bapps): A case with ethiopian students," algorithms, vol. 22, no. 23, p. 24, 2005.